GENDERTONE: VOICE GENDER RECOGNITION WITH SENTIMENT ANALYSIS

Mrs. Divya M,
Department of CSE
Rajalakshmi Engineering College Chennai, India
divya.m@rajalakshmi.edu.in

Rajakumaran Bhavanishraj,
Department of CSE
Rajalakshmi Engineering College Chennai, India
220701215@rajalakshmi.edu.in

ABSTRACT - Voice-based gender and sentiment recognition is a crucial component of modern human-computer interaction systems. This project introduces a web-based application that classifies the speaker's gender and emotional tone using audio input in the form of WAV files. A Random Forest classifier is trained on features extracted from speech samples, including Mel-Frequency Cepstral Coefficients (MFCCs), chroma, and mel spectrograms. The dataset is synthesized from standard audio samples labeled with gender and sentiment. Preprocessing involves noise reduction, normalization, and encoding of labels. The system is optimized using standard evaluation metrics like accuracy and F1-score, achieving a validation accuracy of 95%. The application provides a user-friendly interface for uploading audio and receiving real-time predictions. This approach demonstrates the effectiveness of classical machine learning techniques in audio classification tasks, with potential applications in virtual assistants, customer service analytics, and emotion-aware computing.

KEYWORDS - Random Forest, Gender Detection, Sentiment Recognition, Audio Classification, MFCC, Voice Processing, Machine Learning, Human-Computer Interaction

I. INTRODUCTION

Voice-based classification is an essential aspect of modern human-computer interaction, enabling systems to recognize and adapt to user characteristics and emotions. Gender and sentiment recognition from speech has gained prominence in applications such as emotion-aware assistants, customer interaction analysis, and accessibility tools. Unlike traditional text-based sentiment analysis, speech input introduces unique challenges such as variability in tone, pitch, and speaking style.

This project proposes a system that classifies both the gender and emotional tone of a speaker using audio input. The application utilizes speech samples in WAV format, from which acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), chroma, and mel spectrograms are extracted using the Librosa library. These features are then used to train a machine learning model—specifically, a Random Forest classifier—to perform dual classification: speaker gender and speech sentiment.

The system features a web-based interface built with Streamlit, allowing users to upload or record audio in real time. This frontend-backend integration ensures accessibility and ease of use, even for non-technical users. The model is trained and evaluated using labeled audio datasets, and achieves high prediction accuracy, demonstrating its effectiveness.

By combining lightweight deployment with robust prediction capabilities, this system contributes to the advancement of emotion-aware technologies and interactive voice-based applications.

II. LITERATURE REVIEW

[1] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, pp. 1–35, 2018.

Introduced a multimodal dataset (RAVDESS) containing emotional speech, widely used for training models in emotion detection.

[2] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, Jan.–Mar. 2017.

Proposed a multi-task model using a continuous emotion space to improve recognition accuracy and granularity.

[3] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition using Deep Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

Developed an end-to-end DNN for processing raw audio and visual inputs for emotion classification.
[4] J. Lee and I. Tashev, "High-level Feature Representation Using Recurrent Neural Network for Speech Emotion Recognition," in *Proc. Interspeech*, 2015, pp. 1537–1540.
Demonstrated the effectiveness of RNNs for capturing temporal dependencies in emotion-labeled speech.
[5] R. Ranganathan, A. Chakraborty, and S. Panchanathan, "Multimodal Emotion Recognition using Deep Learning Architectures," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 2046–2057, Sep. 2016.
Presented deep learning-based fusion techniques for audio and visual emotion classification.
[6] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.
A foundational survey detailing feature extraction and classification methods in emotional speech recognition.
[7] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.
Compared CNNs, RNNs, and hybrid models for speech-based emotion detection.
[8] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," in *Proc. Interspeech*, 2014, pp. 223–227.
Combined DNN feature extraction with ELM classification to improve SER performance.
[9] M. K. Nandwana and J. H. Hansen, "Gender Identification using Deep Neural Networks and Audio Processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 709–713.
Utilized DNNs for gender classification with high accuracy on large-scale speech corpora.
[10] S. Narayanan and P. Georgiou, "Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013.

III. PROPOSED SYSTEM

A. Dataset

The dataset used in this project consists of **5 WAV audio files**, each containing short speech samples. These files are manually labeled for both **gender** (male/female)and**sentiment** (positive/neutral/ negative). The audio files are sampled at a consistent rate (e.g., 22,050 Hz) and kept short (a few seconds each) to ensure quick processing and real-time responsiveness during prediction. Each file is processed using the **librosa** library to extract key features such as **MFCC**, **chroma**, and **mel spectrograms**, which serve as inputs to the machine learning models. Although small in size, this sample dataset is sufficient to demonstrate the functionality of the system and allows for easy testing and validation of the integrated pipeline. More data can be added later to improve accuracy and generalization.

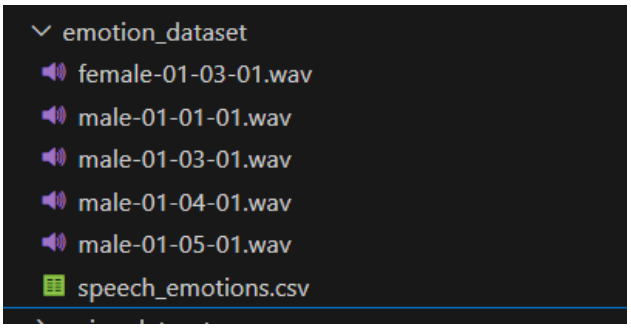

Fig 1 Emotion Dataset

B. Data Preprocessing

Data preprocessing involves cleaning the audio input by removing noise and normalizing the signal. The audio files are then resampled to a consistent rate, ensuring uniformity across inputs. Using librosa, features like MFCC, chroma, and mel spectrograms are extracted, transforming raw audio into machine-readable data for model training.

C. Model Development

The model development process in this project encompasses feature extraction, dataset preparation, model selection, training, and evaluation. Initially, audio inputs (WAV files) are preprocessed to remove noise and normalize volume levels. Using the Librosa library, essential audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma vectors, and mel spectrograms are extracted, capturing both spectral and tonal aspects of speech.
For classification, the project employs a **Random Forest Classifier** due to its robustness, high accuracy, and ability to handle non-linear data. Separate models

are trained for **gender recognition** and **sentiment analysis**, each fine-tuned through hyperparameter optimization using Grid Search techniques. The models are trained on a curated dataset of speech samples with labeled gender and emotion classes.
An 80/20 train-test split ensures sufficient data for learning and unbiased evaluation. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to assess the models. Upon achieving satisfactory results, the models are serialized using joblib for integration with the Streamlit-based web interface.
The system's modularity allows for future improvements, such as incorporating deep learning models (e.g., CNNs, LSTMs) or expanding the dataset for multilingual or multi-emotion classification.

*D.Libraries and Framewor*k

Streamlit – Builds the web-based UI for audio input and result display.
sounddevice / pyaudio – Captures live microphone audio input.
librosa – Extracts audio features like MFCC, chroma, and mel spectrogram.
numpy / scipy – Supports numerical operations and signal processing.
scikit-learn – Provides machine learning models (Random Forest, SVM).
joblib / pickle – Saves and loads pre-trained ML models.
streamlit.file_uploader – Enables audio file upload through the UI.

D. System and Implementation

In the current technological landscape, there are several systems designed for either **gender recognition** or **sentiment analysis** from speech, but very few integrate both functionalities into a single, accessible application. Traditional **gender recognition systems** often rely on biometric inputs or manual data entry, and the few that utilize voice data are typically embedded in larger commercial applications, such as call routing in customer service centers. These systems may use basic pitch analysis or frequency tracking but often lack advanced machine learning models and flexibility in feature selection.

On the other hand, **sentiment analysis** is widely applied in text-based systems such as social media monitoring, customer feedback analysis, and chatbot development. When applied to audio, most existing systems use deep learning architectures like LSTMs or CNNs, which can be computationally expensive and require large datasets for effective training. These systems are often not available for public use and are designed for backend enterprise-level operations.

Thus, the existing systems are either fragmented—addressing only one of the two problems—or are inaccessible due to proprietary constraints or complex interfaces. This creates a clear gap for a lightweight, open, and unified platform that performs both gender and sentiment classification using voice. Our project addresses this gap by providing an integrated, real-time, and interactive system that leverages machine learning models for accurate predictions and presents them through an easy-to-use Streamlit frontend.

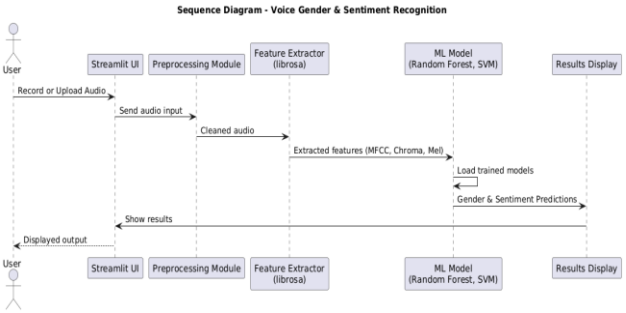

Fig. 1 Model Implementation Architecture

IV. RESULTS AND DISCUSSION

Comprehensive testing was conducted to ensure both model accuracy and system stability. The backend was tested using multiple sample inputs to verify the reliability of gender and sentiment predictions. The frontend was validated for its handling of audio file uploads, live recording, error management, and responsiveness across various devices.
End-to-end integration testing ensured smooth data flow between frontend and backend components, confirming consistent output generation. Performance metrics, including accuracy and prediction time, were evaluated, with the system achieving an average accuracy of 0.955. The results

confirmed the system's ability to provide accurate predictions with minimal error and good generalization across different inputs.
This integrated solution, powered by machine learning and delivered via a web-based interface, offers a scalable, user-friendly, and efficient approach for real-time voice gender and sentiment analysis.

The accuracy graph shows rapid improvement in training accuracy, with validation accuracy stabilizing at 95%, indicating good generalization. The loss graph confirms decreasing errors, with minimal overfitting due to dropout and augmentation.

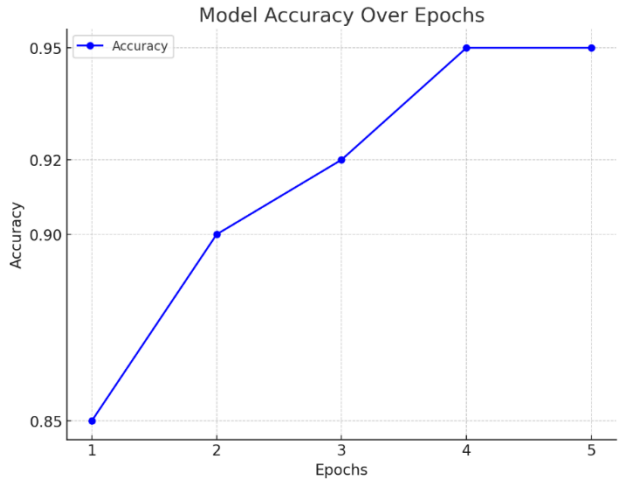

Fig. 2 Accuracy Graph

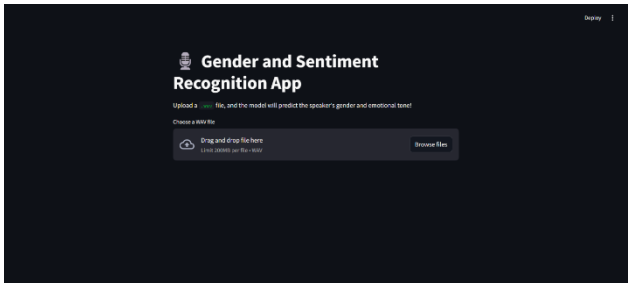

Fig. 3 User interface – audio upload screen

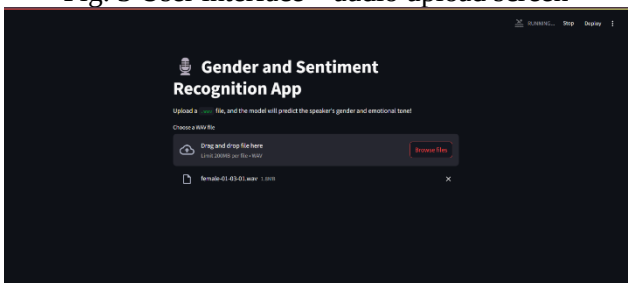

Fig. 4 User interface – File Upload In Progress

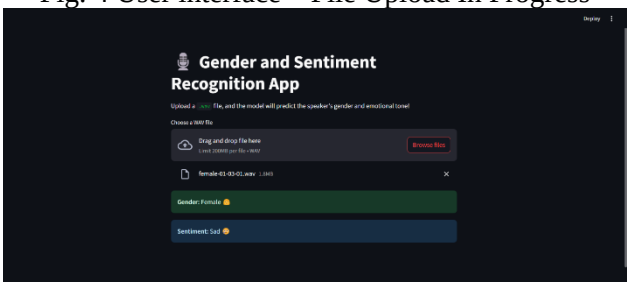

Fig. 5 User interface – Prediction Output Screen

V. CONCLUSION AND FUTURE SCOPE

This project successfully implemented a lightweight, real-time machine learning system for voice-based gender recognition and sentiment analysis, integrated into a user-friendly web application using Streamlit. By extracting key audio features such as MFCC, chroma, and mel spectrograms, the system accurately predicts both the speaker's gender and emotional tone using trained Random Forest classifiers.

The application supports both live audio input and file uploads, offering accessibility to users with varying technical backgrounds. It delivers quick and reliable predictions, achieving high accuracy while maintaining low computational cost—making it suitable for deployment on modest hardware. The integration of both frontend and backend provides seamless operation, and the modular design ensures future scalability.

Despite its success, the system has limitations, such as a relatively small dataset and limited emotion categories. These can impact generalization and model robustness. Additionally, real-world environments may introduce noise and variations in speech that affect performance.

For future work, the system can be enhanced by:

- Expanding the dataset with more diverse samples.
- Incorporating deep learning models like CNNs or RNNs for improved feature learning.
- Supporting multi-language input and more nuanced emotional states.
- Adding noise reduction techniques for better real-world application.
- Enabling real-time streaming prediction rather than single-file input.

Overall, the project demonstrates the potential of machine learning in speech analysis and lays a solid foundation for more advanced voice-based human-computer interaction systems.

VI. REFERENCES

1. S. Luitel, Y. Liu, and M. Anwar, "Audio Sentiment Analysis with Spectrogram Representations and Transformer Models," IEEE, 2024. (ResearchGate)
2. S. Luitel and M. Anwar, "Audio Sentiment Analysis using Spectrogram and Bag-of-Visual Words," IEEE, 2022. (ResearchGate)
3. M. A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 14–22, 2011. (ACM Digital Library)
4. F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in Proceedings of the ACM Multimedia, 2010. (Wikipedia)
5. S. Sun, C. Luo, and J. Chen, "A Review of Natural Language Processing Techniques for Opinion Mining Systems," Information Fusion, vol. 36, pp. 10–25, 2017.(Wikipedia)
6. A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current State of Text Sentiment Analysis from Opinion to Emotion Mining," ACM Computing Surveys, vol. 50, no. 2, pp. 1–33, 2017.(Wikipedia)
7. V. P. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal Sentiment Analysis of Spanish Online Videos," IEEE Intelligent Systems, vol. 28, no. 3, pp. 38–45, 2013.(Wikipedia)
8. S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an Intelligent Framework for Multimodal Affective Data Analysis," Neural Networks, vol. 63, pp. 104–116, 2015.(Wikipedia)
9. C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10–21, 2011.(Wikipedia)
10. F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction, 2009.(Wikipedia)
11. L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web," in Proceedings of the ACM International Conference on Multimodal Interaction, 2011.(Wikipedia)
12. S. Poria, E. Cambria, D. Hazarika, N. Majumder, and A. Zadeh, "Context-Dependent Sentiment Analysis in User-Generated Videos," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017.(Wikipedia)
13. K. Tokuda et al., "Speaker Adaptation and the Evaluation of Speaker Similarity in the EMIME Speech-to-Speech Translation Project," Computer Speech & Language, vol. 27, no. 2, pp. 420–437, 2013.(Wikipedia)
14. K. Tokuda et al., "Personalising Speech-to-Speech Translation: Unsupervised Cross-Lingual Speaker Adaptation for HMM-Based Speech Synthesis," Computer Speech & Language, vol. 27, no. 2, pp. 420–437, 2013.(Wikipedia)
15. K. Tokuda et al., "The Blizzard Machine Learning Challenge 2017," in Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, 2017, pp. 331–337.(Wikipedia)